用Excel绘制统计图

这一节虽然名字被称为算法，实际上更准确的应该换成方法。因为我们这一节主要使用到的软件是Excel，我们将介绍在Excel中绘制统计图的方法。

事实上，统计图最常用的就几种：饼图、柱状图、条形图、直方图、折线图 、散点图、箱线图等。对于同一组数据，不同的同学绘制出来的效果可能截然不同，好的图形能够让读者一样就能看出数据的规律和特点，其传递给读者的信息是准确和有效的，准确是指我们不能随意绘制图形，因为每个统计图适用的数据是有限制的；有效是指图形传递出来的信息和你得到的结论是吻合的；这是一个图形要满足的最基本两点。另外，要评价一个图形是否画的好，主要看其是否简洁和美观，我们绘制出来的图表包含的信息一定要清晰明显，不能花里胡哨，另外一定要注意图形的配色。

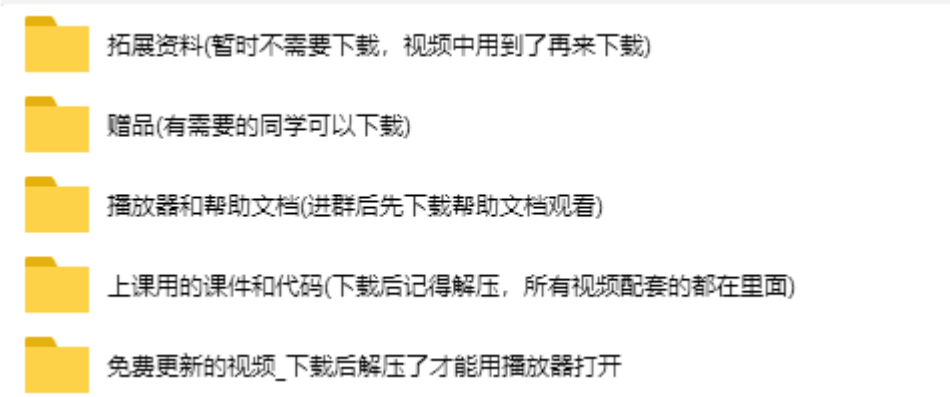
注意：我使用的版本为Office2016，较低版本的Office或WPS可能没有视频中涉及到的一些功能或选项，大家可以根据自己的软件版本灵活处理，我也比较推荐大家用Office2016或2019，版本越高内置的功能越丰富，需要安装新版本的同学可以去关注我的微信公众号“数学建模学习交流”后发送软件两个字。

（注：本讲所有数据都是随机生成的）

温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、我视频中推荐的资料**等。



（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

数学建模学习交流

饼图

最适合采用饼图的情形：

1. 只有一个数据系列（单分类数据）。
2. 任何数据值都不为零或小于零。
3. 类别不超过七个。因为七个以上的扇区会使图表难以阅读。
4. 划分的类别最好是完整的，一般不完整时可以加其他。
5. 类别过多可用复合饼图时，千万别硬要画一个完整饼图。
6. 类别只有两个时就不用画图了，没多大必要。

注意：不用在图中加入标题，我们一般在论文的正文中加入（表上图下）。另外，画出来的图一定要有分析，要告诉读者你画图的目的是什么。

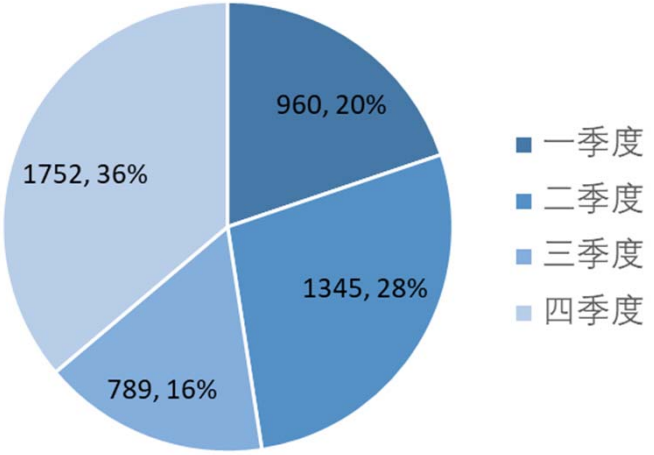


图3.1 某企业2018年四季度女装销量（万件）

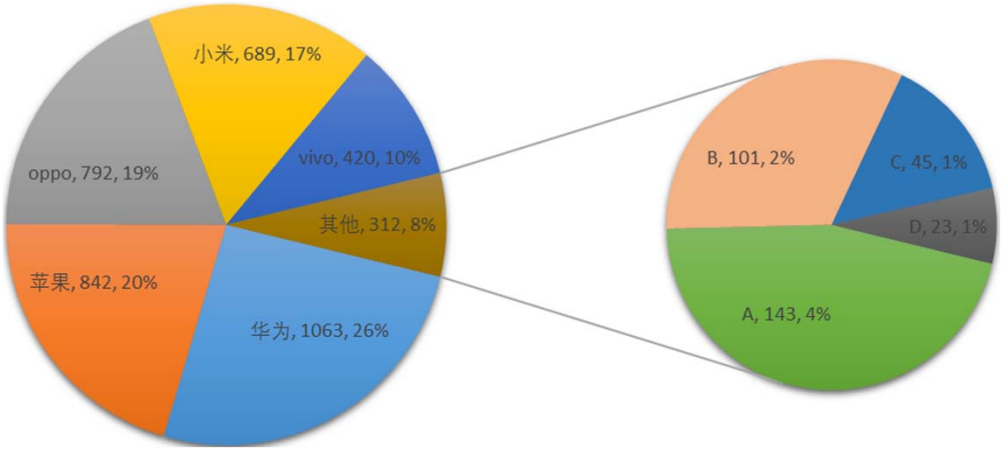
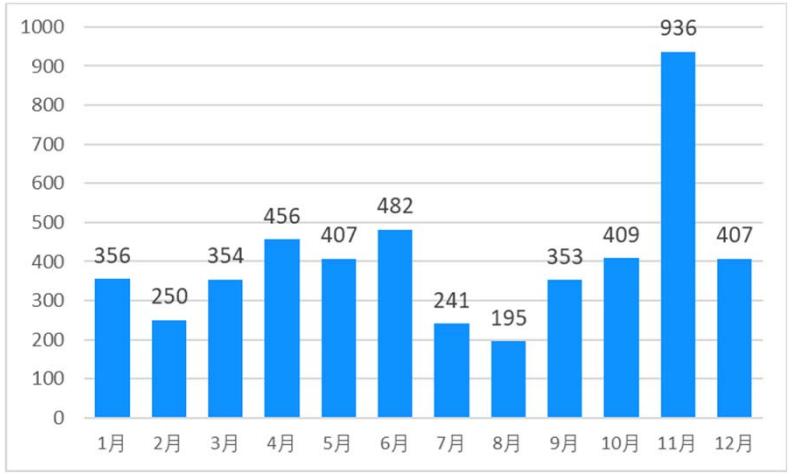
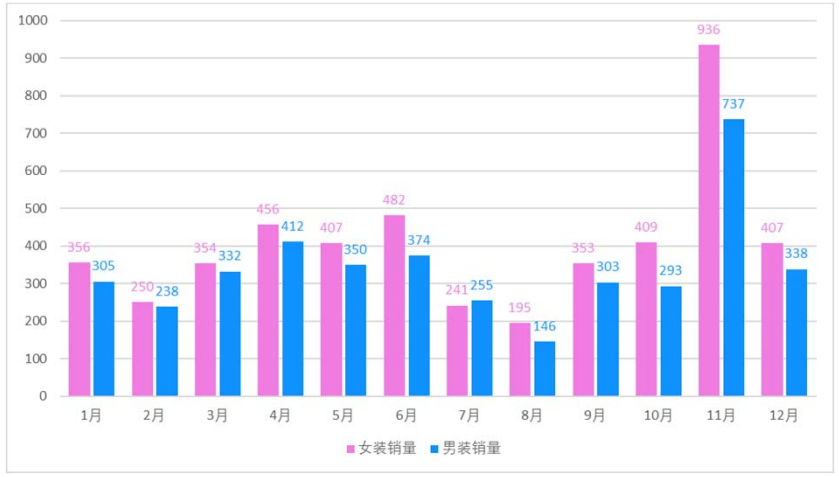


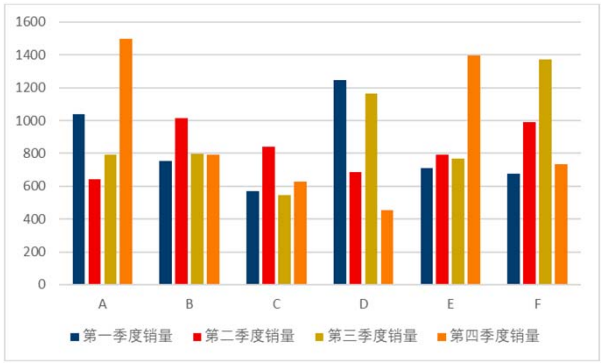
图3.1 中国2016年各品牌手机销量（万台）

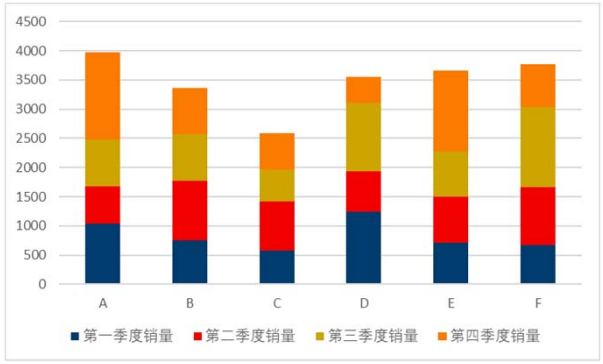
柱状（形）图

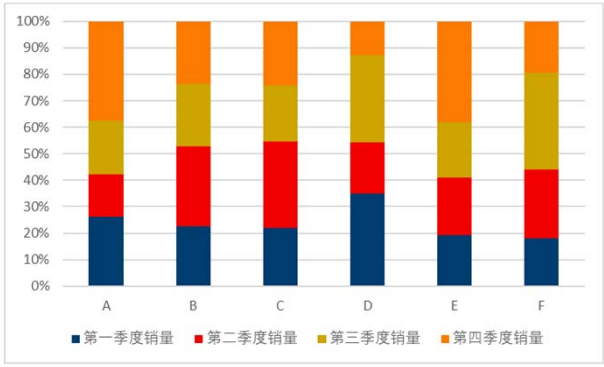
限于篇幅，下面的图都比较小，高清大图可在Excel文件中查看。





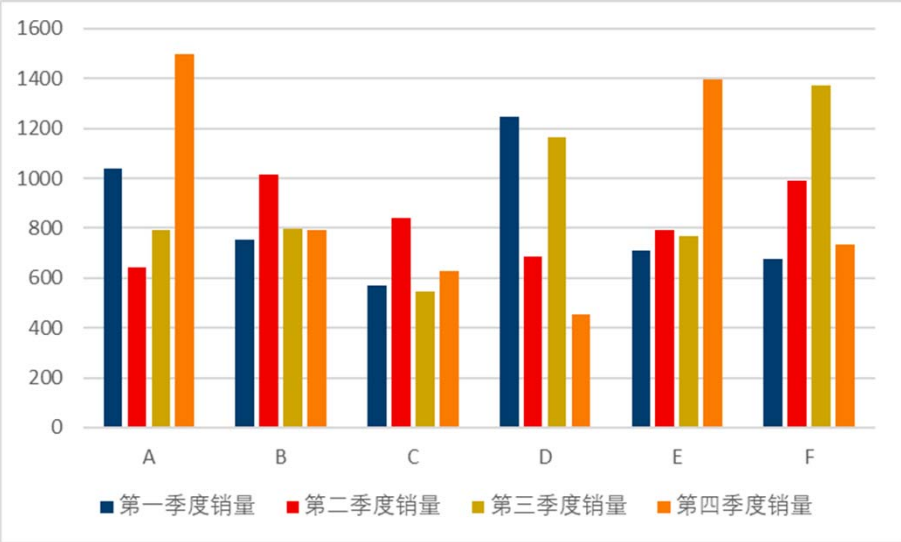


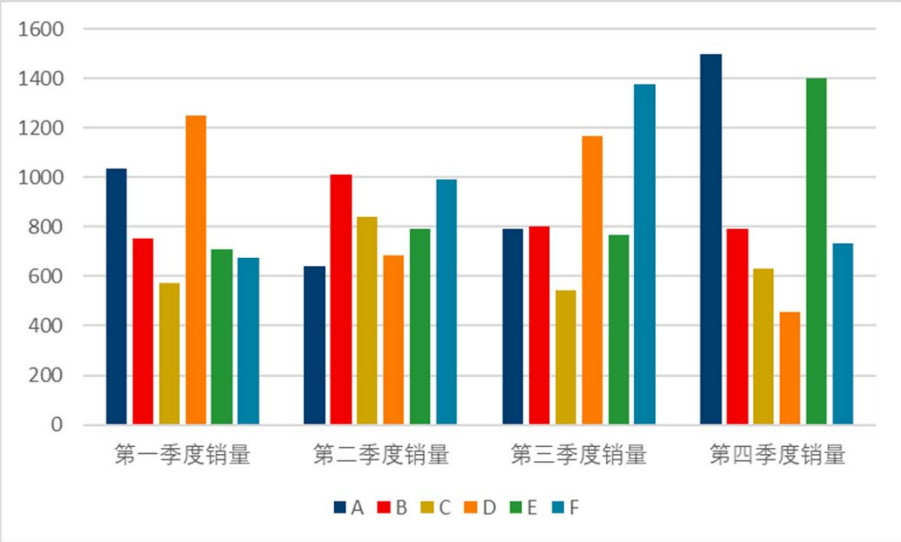


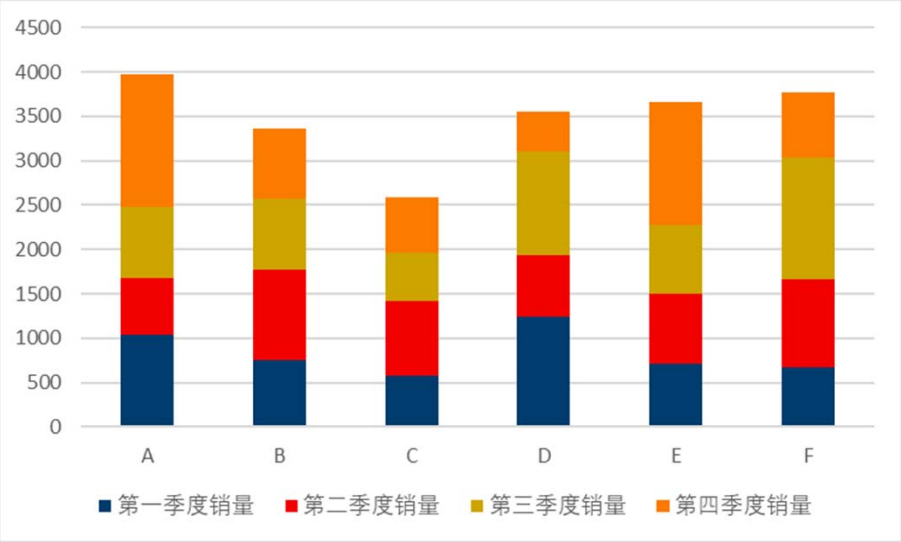


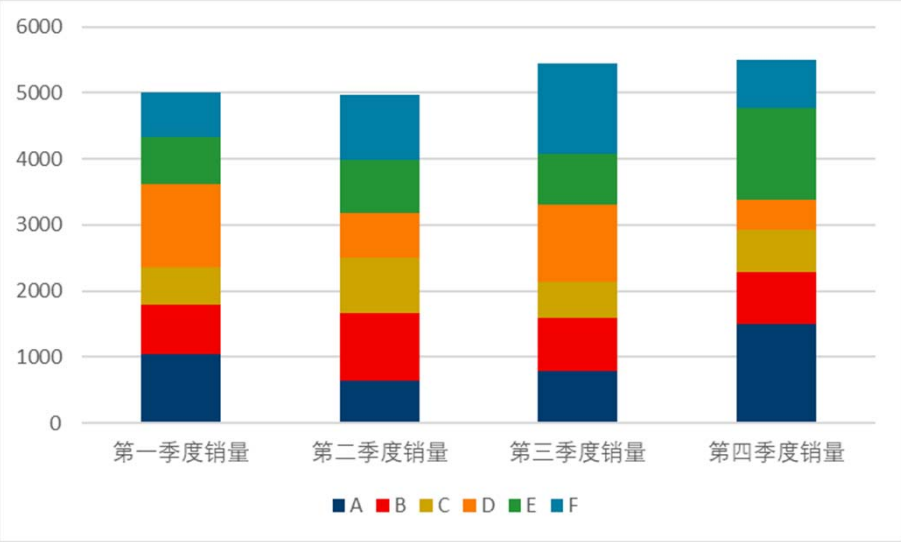
柱状图常常用于展示多个分类（单个分类也可以）的数据变化和同类别各变量之间的比较情况。堆积柱状图可用于比较同类别各变量和不同类别变量总和差异；百分比堆积柱状图适合展示同类别的每个变量的比例。

切换行和列后的柱状图（左→右）

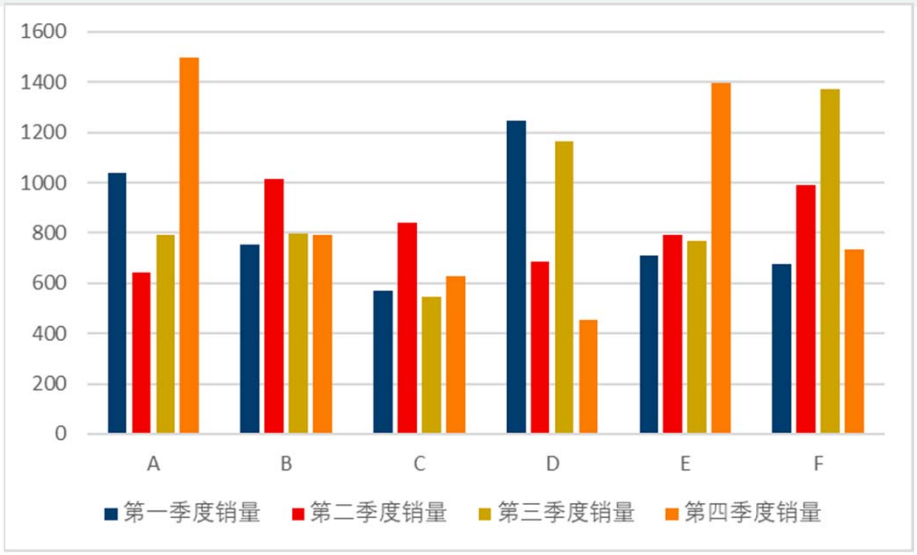


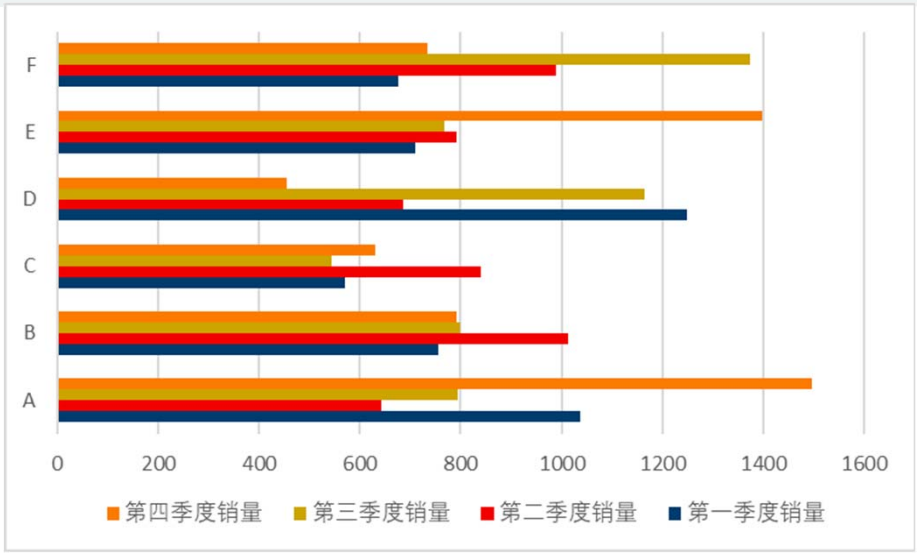


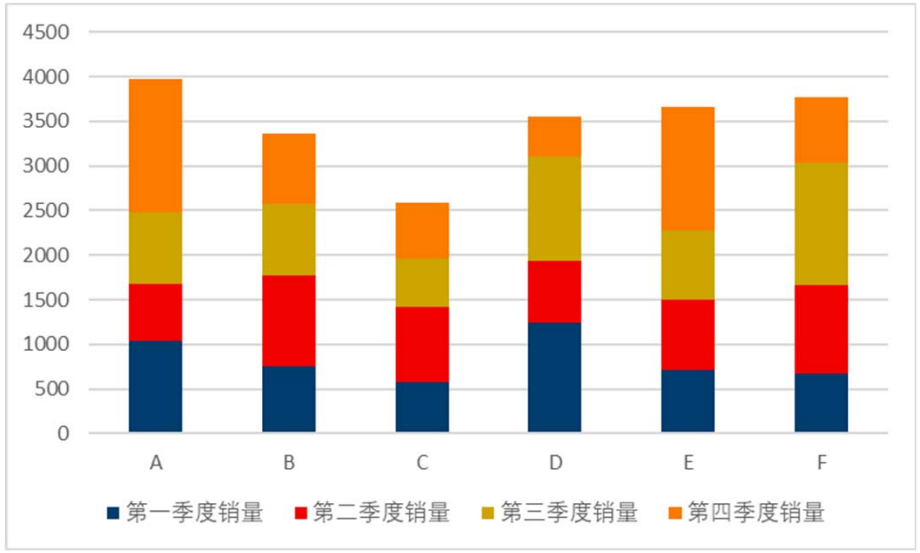


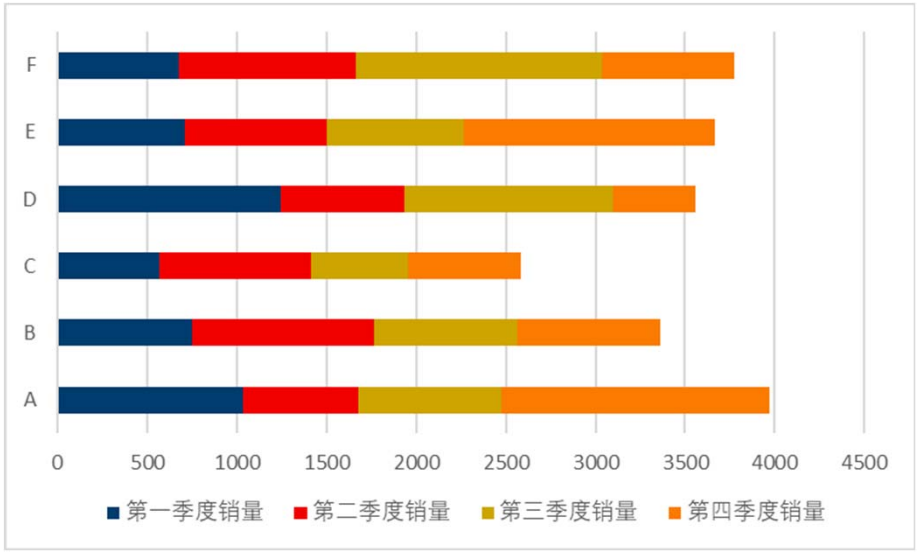


条形图（横过来的柱状图）



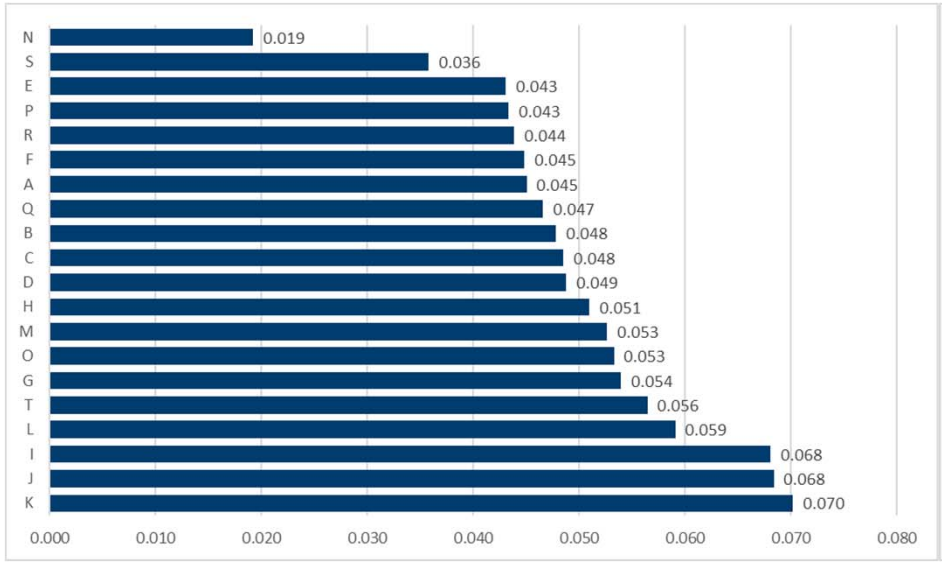




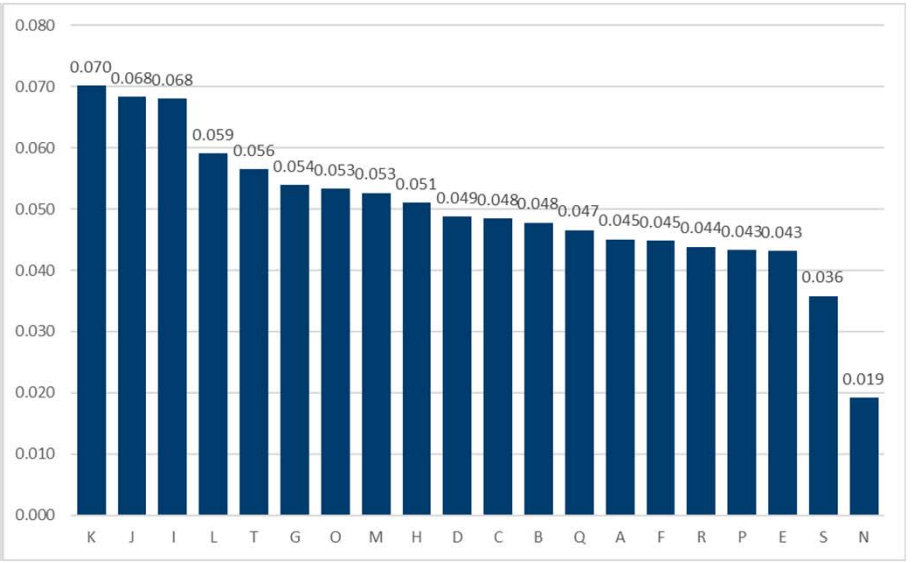


条形图的优势

这是我们第二讲介绍topsis评价模型时得到的二十条河流的评分图，原始数据已经经过了排序。



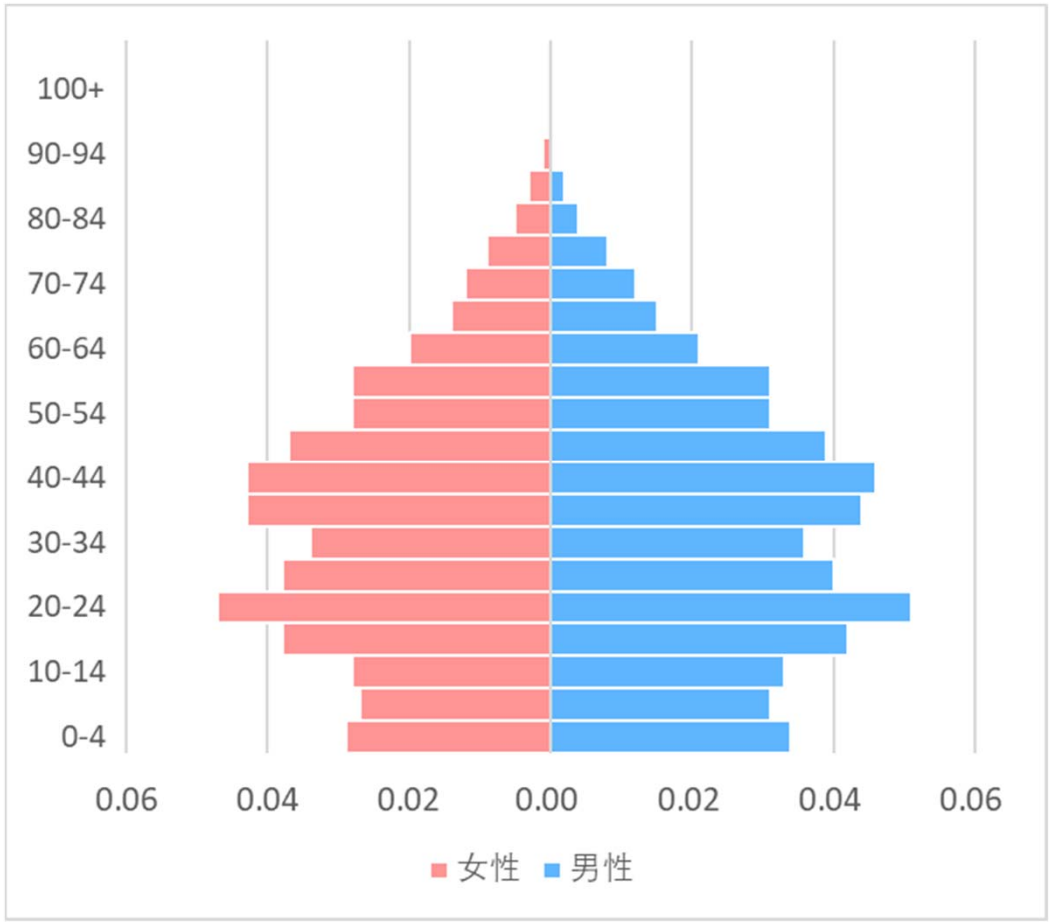
条形图（推荐）



柱状图

类别数过多时，如果要加入数据标签，那么使用条形图比较合适
柱状图的数看起来有点拥挤

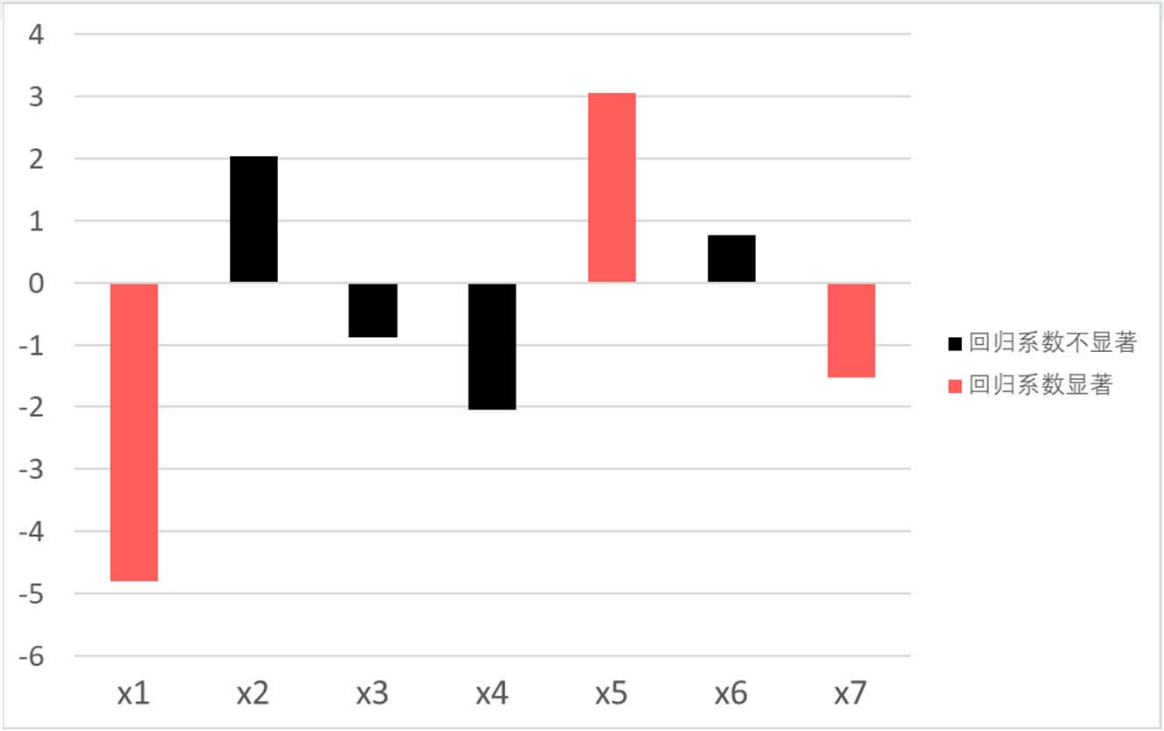
双向条形图



2010中国人口结构图（年龄金字塔图）

用柱状图可视化回归结果

自变量	回归系数	P值(显著性水平取0.05)
x1	-4.80	0.0289
x2	2.04	0.0789
x3	-0.86	0.1259
x4	-2.05	0.2365
x5	3.06	0.0001
x6	0.76	0.4253
x7	-1.54	0.0403



对图形的解读：

（1）用红色和黑色区分了显著和不显著的系数估计。红色是指系数估计跟0有显著差异，而黑色是指没有。因此解读的时候，关注红色柱子即可。

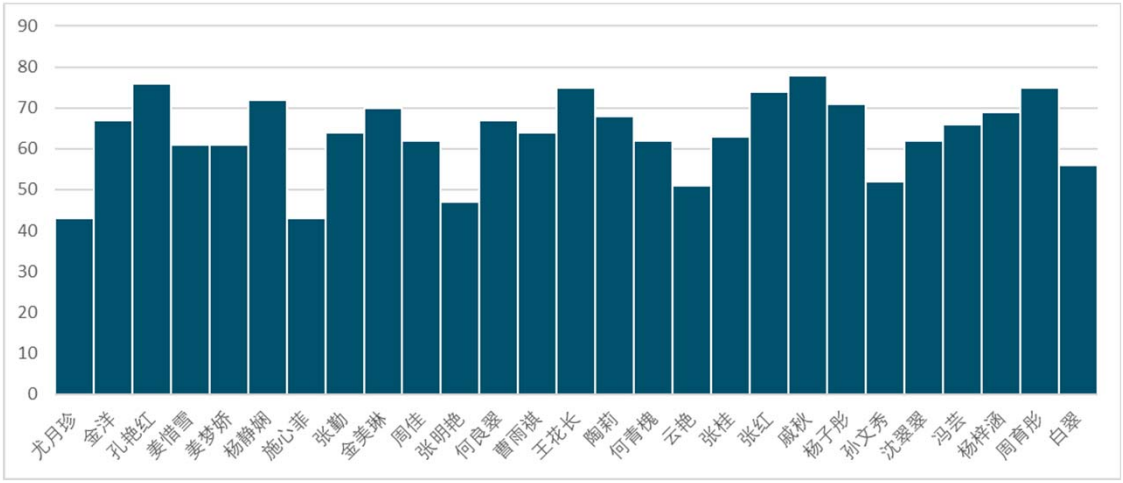
（2）柱子朝上，说明自变量和因变量的关系是正向的。自变量取值增加的时候，因变量取值也增加。类似地，如果柱子朝下，说明自变量和因变量的关系是负向的。自变量取值越大，因变量取值则越小。

（3）若对自变量进行了标准化，那么柱子的高度，也就是系数的估计值有可比性，可以直观地区分出自变量对因变量的影响大小。

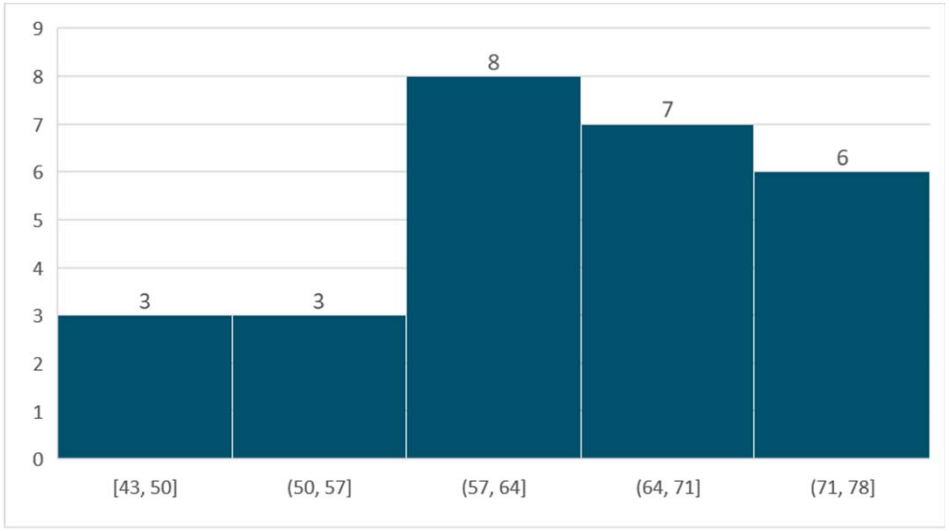
直方图

很多同学区分不开直方图和柱状图，事实上：**直方图是显示数据频数或频率的柱状图。**

姓名	体重
尤月珍	43
金洋	67
孔艳红	76
姜惜雪	61
姜梦娇	61
杨静娴	72
施心菲	43
张勤	64
金美琳	70
周佳	62
张明艳	47
何良翠	67
曹雨祺	64
王花长	75
陶莉	68
何青槐	62
云艳	51
张桂	63
张红	74
戚秋	78
杨子彤	71
孙文秀	52
沈翠翠	62
冯芸	66
杨梓涵	69
周育彤	75
白翠	56



柱状图



直方图

直方图和柱状图的区别

直方图（Histogram）是一种可视化在连续间隔，或者是特定时间段内数据分布情况的图表，经常被用在统计学领域。简单来说，直方图描述的是一组数据的频次分布，例如把年龄分成“0-5,5-10，……，80-85”17个组，统计一下中国人口年龄的分布情况。直方图有助于我们知道数据的分布情况，诸如众数、中位数的大致位置、数据是否存在缺口或者异常值。
（注：众数是指一组数据中出现次数最多的数据值，众数可能是一个数，但也可能是多个数。中位数是指可将数值集合划分为相等的上下两部分的数。）

直方图和柱状图最让人迷惑的地方，就是它们长得非常相似。实际上，直方图和柱状图无论是在图表意义、适用数据上，还是图表绘制上，都有很大的不同。

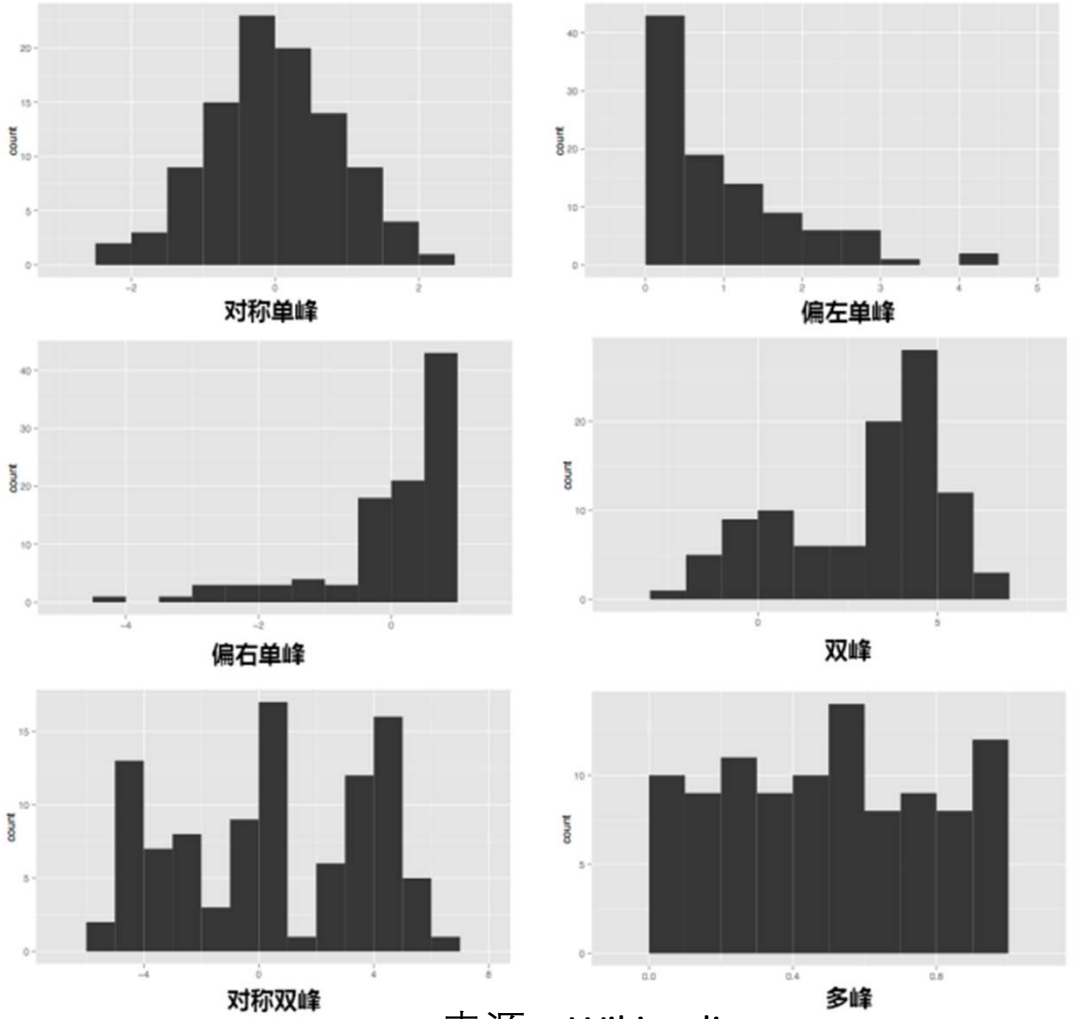
1.直方图展示数据的分布，柱状图比较数据的大小。

2.直方图X轴为定量数据，柱状图X轴为分类数据。

3.直方图y轴要么为数据的频数，要么为数据的频率，柱状图y轴为数据实际大小。

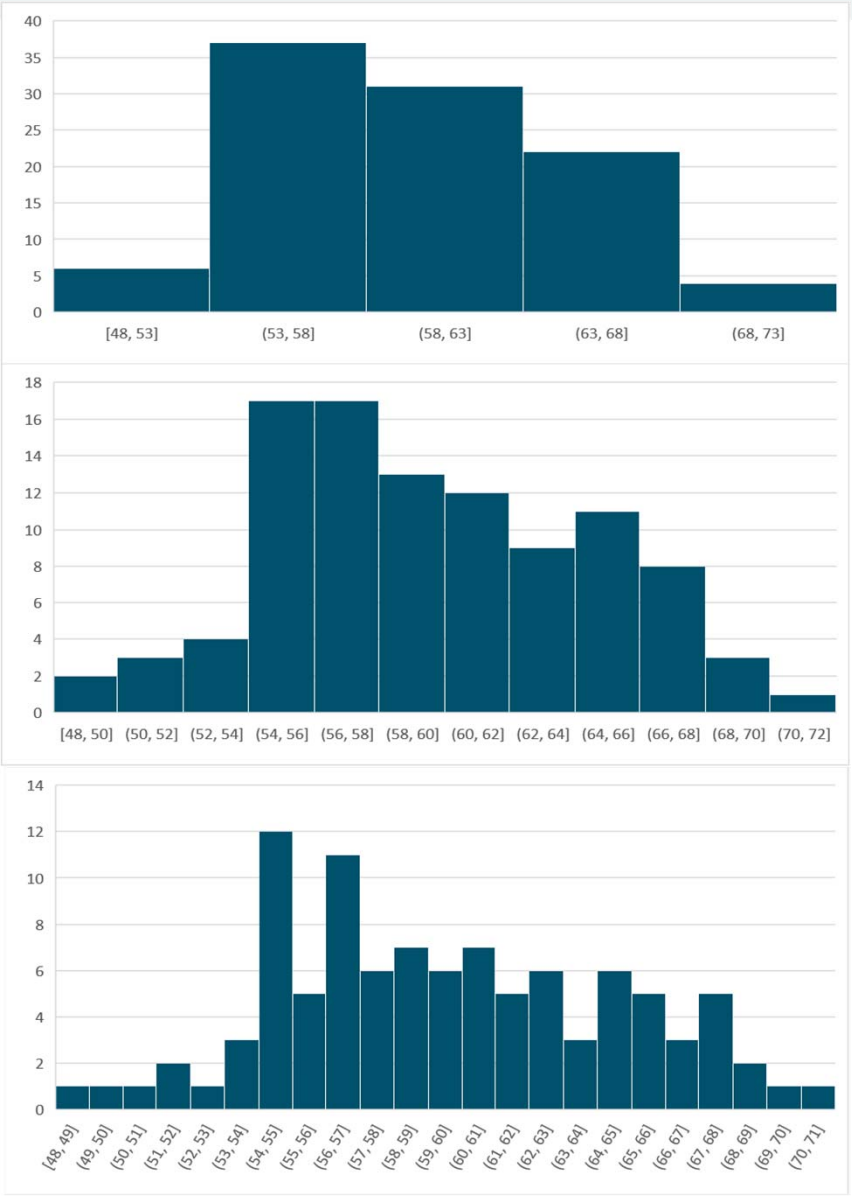
直方图和柱状图的区别

根据数据分布状况不同，直方图展示的数据有不同的模式，常见的如下所示：



来源：Wikipedia

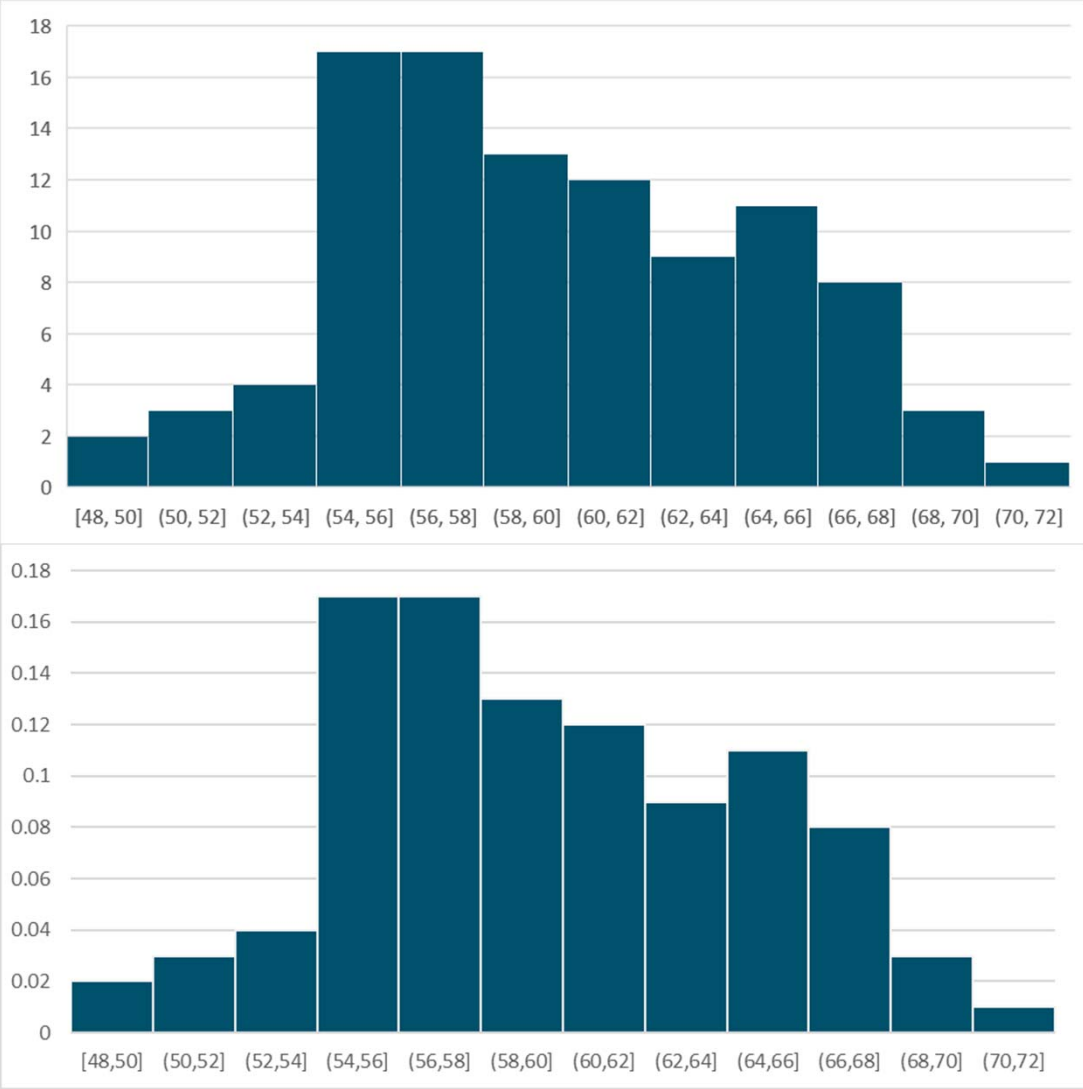
组距对于直方图的影响



组距会影响直方图呈现出来的数据分布，因此在绘制直方图的时候需要多次尝试改变组距。

左图从上到下的组距分别为：5, 2, 1.

频数和频率分布直方图

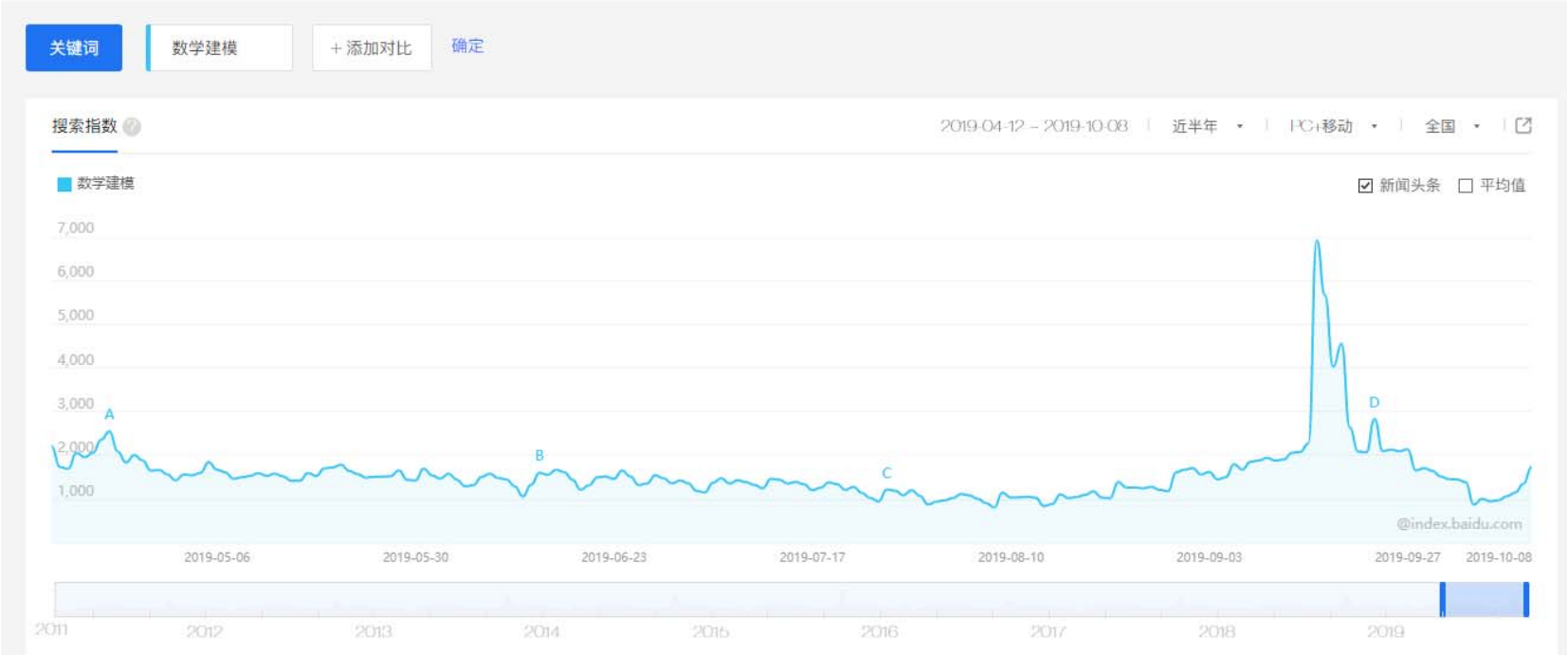


注意：有的地方的频率分布直方图的纵坐标取的是频率/组距，大多数情况下直方图对应的各个类别组距相等，因此得到的图形和我们这里没有实质的区别，仅仅相差了一个倍数关系。

折线图

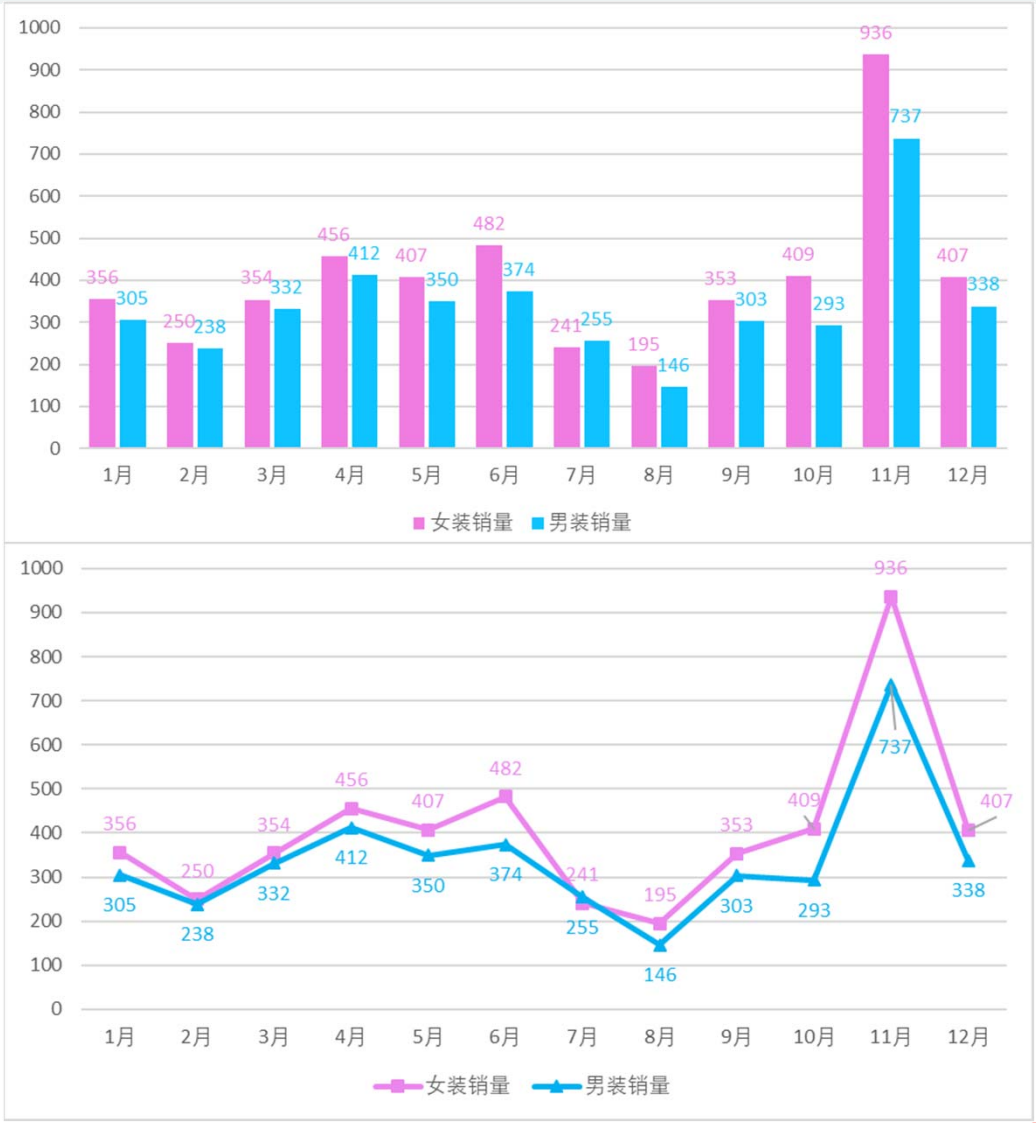
如果数据是时间序列数据（如日、月、季度或年度数据），则应该考虑使用折线图，尤其是时间跨度长且存在多个时间序列时，更应该使用折线图 。（折线图也经常被称为时间序列图，或简称为时序图）

当然，时间序列期数较少时，也可以考虑使用柱状图哦。



图：百度指数搜索数学建模得到的折线图

折线图和柱状图的对比

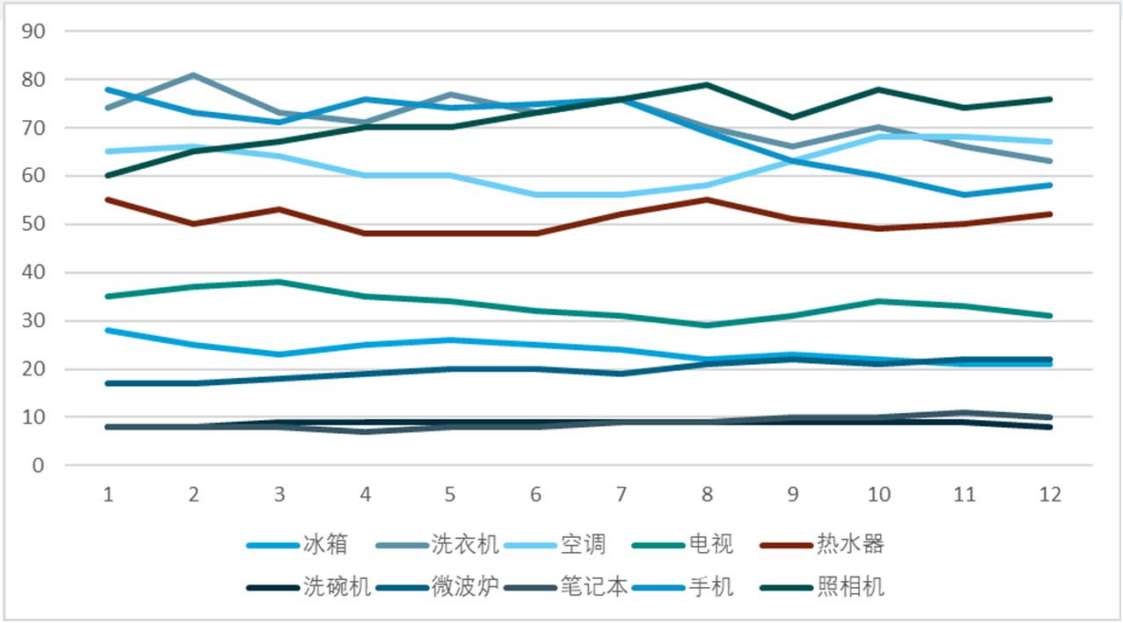


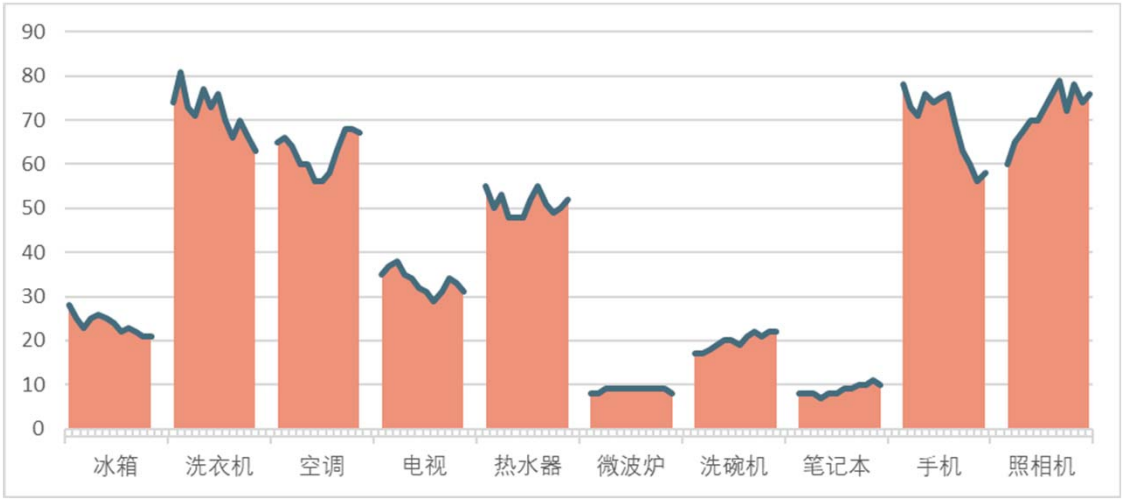
双坐标轴折线图

2016年至今交易日的上证指数

2016年至今交易日的上证指数及其增长率

柱形图顶端的折线图

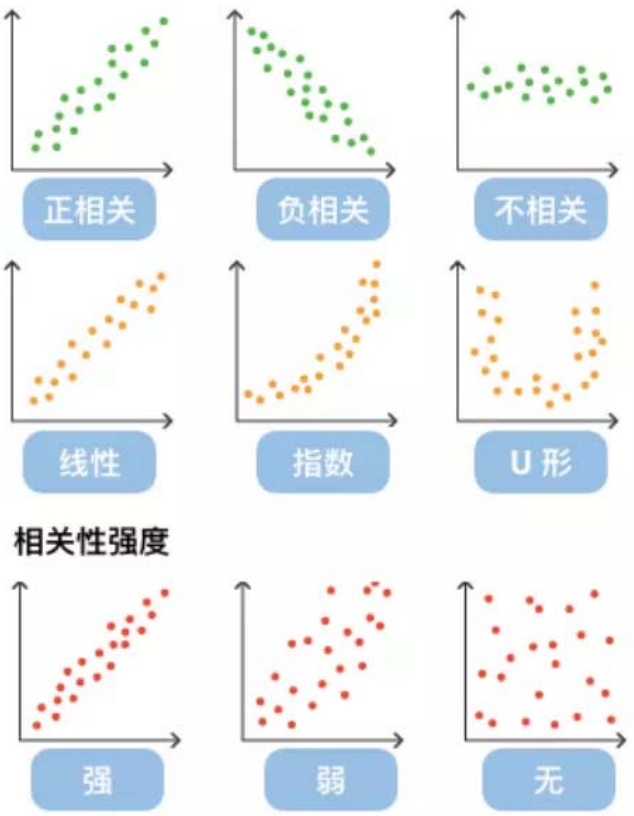




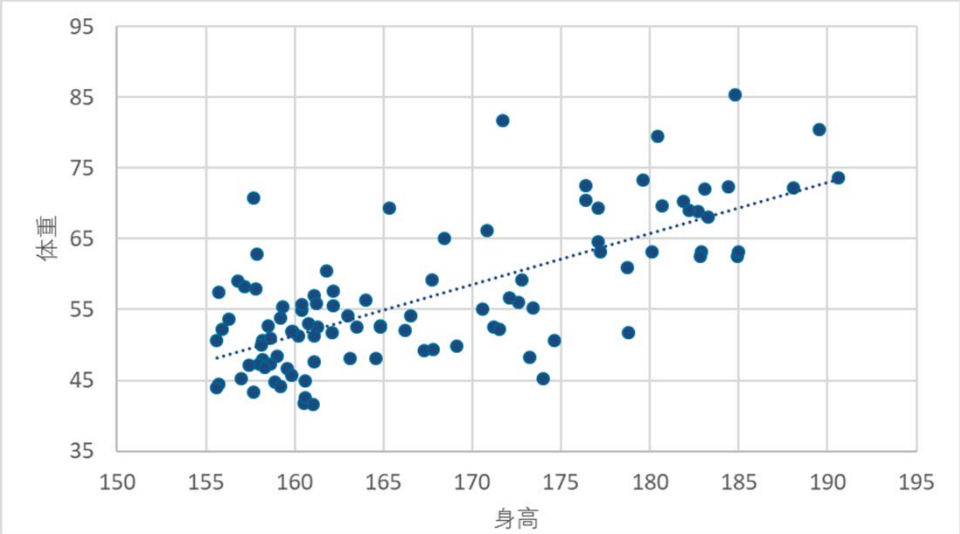
散点图

散点图也叫 X-Y 图，它将所有的数据以点的形式展现在直角坐标系上，以显示变量之间的相互影响程度，点的位置由变量的数值决定。

通过观察散点图上数据点的分布情况，我们可以推断出变量间的相关性。如果变量之间不存在相互关系，那么在散点图上就会表现为随机分布的离散的点，如果存在某种相关性，那么大部分的数据点就会相对密集并以某种趋势呈现。数据的相关关系主要分为：正相关（两个变量值同时增长）、负相关（一个变量值增加另一个变量值下降）、不相关、线性相关、指数相关等，表现在散点图上的大致分布如右图所示。那些离点集群较远的点我们称为离群点或者异常点。

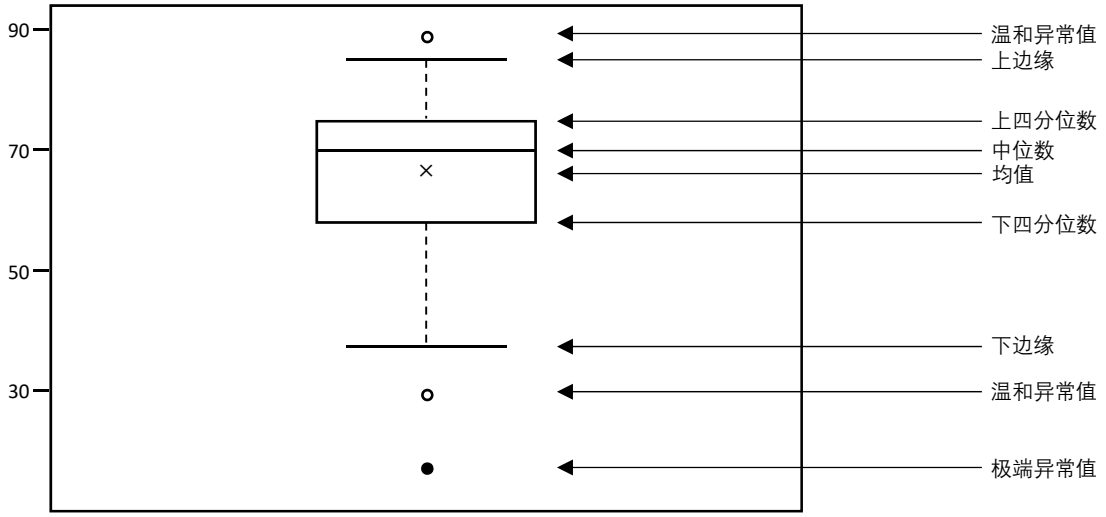


带标识的散点图



箱线图

箱线图也称箱须图、箱形图、盒图，用于反映一组或多组连续型定量数据分布的中心位置和散布范围。箱形图包含数学统计量，不仅能够分析不同类别数据各层次水平差异，还能揭示数据间离散程度、异常值、分布差异等等。



计算过程（注意：箱线图有不同的画法，下面介绍的是用的较多的一种画法）：

1 计算上四分位数、中位数、下四分位数以及均值；

2 计算上四分位数和下四分位数之间的差值，即四分位数差（IQR，interquartile range）；

3 绘制箱线图的上下范围，上限为上四分位数，下限为下四分位数，在箱子内部中位数的位置绘制横线；

4 大于上四分位数1.5倍四分位数差的值，或者小于下四分位数1.5倍四分位数差的值，划为异常值（outliers）；

5 排除掉异常值之外，在剩下的数据的最大值和最小值处画横线，作为箱线图的上下边缘；

6 极端异常值，即超出四分位数差3倍距离的异常值，用实心点表示；较为温和的异常值，即处于1.5倍-3倍四分位数差之间的异常值，用空心点表示；

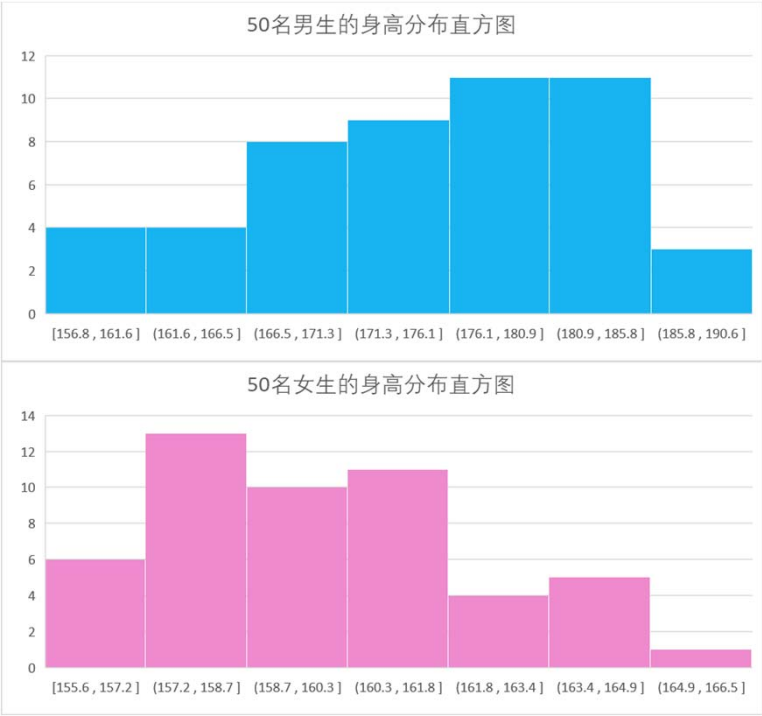
7 为箱线图添加名称，数轴等，并在图中用×标记出数据的均值位置。

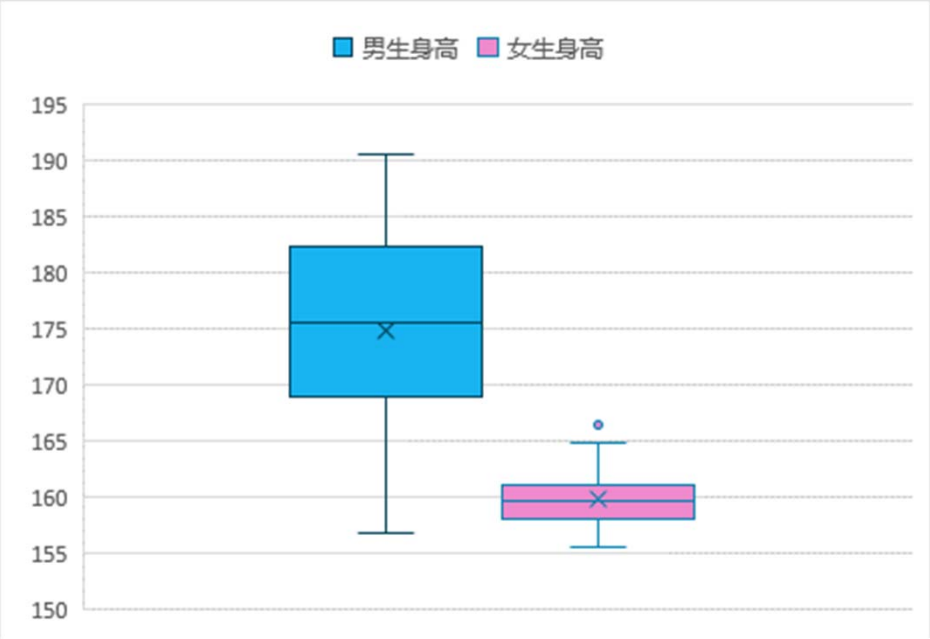
箱线图的作用

箱线图的用法是，配合定性变量画分组箱线图，作比较。如果只有一个定量变量，很少用一个箱线图去展示其分布，更多选择直方图。箱线图更有效的使用方法是作比较。

假设要比较男女生的身高，用什么工具最好？答案是箱线图。

箱线图明显更加有效，能够从平均水平（中位数）、波动程度（箱子高度）以及异常值对男女教师的教学评估得分进行比较，而直方图却做不到。





注意：[2 3 3 5 6 8 8 9 9 10]的下四分位数是3，视频中口误了